# Improved Mixing in MCMC Algorithms for Linear Models

**Todd L. Graves**

Los Alamos National Laboratory, PO Box 1663, MS F600, Los Alamos, NM, 87545

**Paul L. Speckman and Dongchu Sun**

Department of Statistics, University of Missouri-Columbia, Columbia, MO 65210

## Abstract

In spite of increasing computing power, many statistical problems give rise to Markov chain Monte Carlo algorithms that mix too slowly to be useful. Often, the mixing is due to high posterior correlations between parameters. In the illustrative special case of Gaussian linear models with known variances, we characterize which functions of parameters mix poorly and demonstrate a practical way of greatly reducing autocorrelations from the algorithms by adding Gibbs steps in suitable directions. These additional steps, and their Metropolis analogues, are also very effective in practical problems with nonnormal posteriors and are particularly easy to implement.

# 1 Introduction

Following Gelfand & Smith (1990), MCMC methods have become important and popular, especially for Bayesian computation. Although the MCMC method will converge in a very general setting (e.g. Tierney 1994), it is well known that in many applications of raw MCMC methods, in particular the Gibbs sampler or Metropolis-Hastings sampling, successive parameter updates tend to be highly correlated, so convergence in MCMC to the stationary posterior distribution is painfully slow. When the models become more complicated, it becomes increasingly likely that untuned methods will not mix quickly enough to be practical.

Many of the issues of convergence of MCMC chains have been discussed in Gilks et al. (1996). See especially Gilks & Roberts (1996). There are several possible solutions. One is to run MCMC longer, using tools such as those of Gelman & Rubin (1992) and Gelman (1996) to monitor convergence. Another method is to use block sampling. See, for example, Liu et al. (1994) and Roberts & Sahu (1997) or Chib & Carlin (1999). Other proposals involve reparameterizing, adding parameters, or both. The sweeping method of Vines et al. (1996) and hierarchical centering of Gelfand et al. (1995) and Gelfand et al. (1996) both deal with essentially overparameterized linear or linearizable models. An interesting variant on the latter is the partial centered parameterizations of Papaspiliopoulos et al. (2003). In contexts where reparameterization appears inadequate, data augmentation and parameter expansion are sometimes helpful, for example, as in parameter-expanded data augmentation (Liu & Wu (1999), Meng & van Dyk (1999). Parameters added to a model may reduce *a posteriori* correlation and improve mixing. Reparameterization for computatational efficiency has a long history; Gelman (2004) shows how reparameterization may even add statistical insight.

In contrast to these approaches, we discuss a strategy for enhancing mixing in standard MCMC algorithms for standard linearizable models and their natural or customary priors. The strategy is based on hybrid MCMC algorithms that augment standard cycles with extra moves in well-chosen directions in the parameter space without adding parameters. We call these *decorrelation steps* because they tend to enhance mixing and reduce, sometimes dramatically, the correlation between successive samples from the posterior. The puspose of this paper is to demonstrate the effectiveness of these hybrid algorithms and their ease of implementation. Our main theoretical results pertain to Gaussian posteriors, where we show that decorrelation steps in some common situations can nearly eliminate autocorrelation between successive cycles in MCMC chains. Our work is anticipated in the hybrid sampler of Nobile (1998), who introduced additional scale moves to improve mixing in certain multinomial problems. Liu & Sabatti (1999) used mixed strategies including extra sadditive steps as used here, and we rely on the work of Liu & Sabatti (2000) to verify that decorrelation steps preserve the posterior distribution.

To motivate the discussion, consider one of the simplest possible linear models, one-way analysis of variance. This example is discussed by virtually every author concerned with mixing.

**Example 1: One-way ANOVA** Let

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, m; j = 1, \ldots, n, \tag{1}$$

where the $\varepsilon_{ij}$ are independent $N(0, \delta_0)$, with conventional independent priors $\mu \sim N(0, \delta_1)$ and $\alpha_i \overset{iid}{\sim} N(0, \delta_2)$. (To complete a hierarchical specification, conjugate inverse-gamma priors can be taken on the $\delta_i$.) In the fixed variance case, this model is known to display slow convergence under conditions on the relative values of the variances (see, for example, Gelfand et al. (1995)). The reason seems intuitively clear; the model is poorly specified in that the cell means $\mu_i = \mu + \alpha_i$ do not uniquely determine the parameters $\mu$ and $\alpha_i$. In classical linear models, statisticians only perform inference on estimable functions of the parameters. In the Bayesian setup, with proper priors, the posterior is proper even if it is relatively noninformative in certain directions, so a statistician may desire to use the model even though it is poorly identifiable.

One simple solution to the problem of poor mixing is reparameterization. The *sweeping* method of Vines et al. (1996) replaces $\alpha_i$ by $\tilde{\alpha}_i = \alpha_i - \bar{\alpha}_i$ with corresponding change in the prior. This strategy corresponds to the standard frequentist method of adding one or more side conditions to estimate a nonidentifiable model. Sweeping trades simplicity of structure of the model for efficient mixing of the MCMC algorithm. Vines et al. (1996) (see also Gilks et al. 1996) show how sweeping can be extended to higher order models. Of course, in non-Gaussian settings, full conditional distributions after transformation may no longer have closed form, complicating MCMC algorithms.

**Example 2: Hierarchical One-way ANOVA** Another popular reparameterization is the cell mean model, referred to as hierarchical centering by Gelfand et al. (1995), where

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, m; j = 1, \ldots, n, \tag{2}$$

with $\varepsilon_{ij}$ as above but independent $\mu_i \sim N(\mu, \delta_2)$ with a normal prior on $\mu$ and again conjugate priors on the variance components. Gelfand et al. (1995) showed that if $n \to \infty$, the posterior correlation of $\alpha_i$ and $\alpha_j$ in model (1) will go to 1, while that of $\mu_i$ and $\mu_j$ will go to 0. Consequently, the MCMC chain based on model (2) is much more efficient. Hierarchical centering works extremely well in the one-way ANOVA model or a nested higher way ANOVA if the data variance is low (e.g., if $m$ is large enough). When combined with sweeping, hierarchical centering is perhaps the most popular strategy for efficiently obtaining posterior quantities using MCMC. In Section 3.4, we show why sweeping must be added to hierarchical centering in two-way or multi-way ANOVA models, with or without interactions.

Papaspiliopoulos et al. (2003) have proposed an interesting variant of (2) called a partially noncentered parameterization. While the centered model (2) is effective when the information from the data outweighs that from the prior, it is poor for the high data variance case where (1) has better *a posteriori* correlation. To treat both cases with one model, Papaspiliopoulos (2003) and Papaspiliopoulos et al. (2003) proposed a continuum of *partially noncentered parameterizations*. For the illustrative case of one-way ANOVA, the model is

$$\begin{aligned} y_{ij} &= w\mu + \tilde{\beta}_i^w + \varepsilon_{ij}, j = 1, \ldots, n, \\ \tilde{\beta}_i^w &= (1-w)\mu + z_i, i = 1, \ldots, m, \end{aligned} \tag{3}$$

where $\mu \sim \mathrm{N}(0, \delta_1)$ is independent of $z_i \overset{iid}{\sim} N(0, \delta_2)$ and $w$ is a fixed number in $[0, 1]$. Taking $w = 1$ gives (1), while $w = 0$ yields (2). Further discussion is given in Section 3.4.3.

In implementing a general purpose Bayesian computation package called YADAS (Graves (2003*a*)), one of the authors has found that periodic extra MCMC steps in well-chosen directions appear to dramatically improve convergence and result in greatly reduced correlation in the MCMC parameter updates. For example, in model (1), a Metropolis-Hastings step with proposal $(\mu + Z, \alpha_1 - Z, \ldots, \alpha_m - Z)$, where $Z$ is an independent sample from a convenient symmetric distribution, serves the purpose well. See Graves (2003*b*), and see also Graves et al. (2003) and Graves & Picard (2003) for other motivating examples.

The central idea of this paper is that problems with convergence in MCMC sometimes arise because the model is poorly specified, as in model (1). Even though a proper prior may be specified and the posterior is proper, the lack of identifiability in the basic linear structure of the likelihood leads to high correlation among MCMC updates. In such cases, a well-chosen step in a new direction can be beneficial.

We focus on the case of linear models or derivatives. Such models are characterized by the fact that the likelihood depends on a parameter vector $\boldsymbol{\beta}$ only through values of $\boldsymbol{X\beta}$, for some design matrix $\boldsymbol{X}$. Obvious examples include generalized linear models with link $g(\boldsymbol{\mu}) = \boldsymbol{X\beta}$, although other distributional setups are possible. Let $\mathcal{S} = \mathcal{N}(\boldsymbol{X'X})$ denote the null space of $\boldsymbol{X'X}$. If the prior on $\mathcal{S}$ is flat so the posterior is improper, Besag et al. (1995) and Gelfand & Sahu (1999) have shown how adjustments in the null space to recenter the Markov chain induce convergence to a stationary chain with the desired posterior distribution. We extend the idea for use with proper posteriors. Ignoring other parameters for the moment, suppose a prior on $\boldsymbol{\beta}$ is chosen with proper posterior, and MCMC is used to estimate *a posteriori* quantities. Denoting the output from chain as $\boldsymbol{\beta}^{(k)}$, we suppose high autocorrelation between successive samples. The method of Graves and his coauthors is to augment the standard Markov chain cycle with one or more additional Metropolis-Hastings steps having proposals of the form $\boldsymbol{\beta}+\boldsymbol{g}Z$ for $\boldsymbol{g} \in \mathcal{S}$ and symmetrically distributed $Z$. Intuitively, these moves tend to shift the Markov chain toward (or directly away from) the center of the posterior distribution, thereby interrupting the random walk nature that typifies chains generated from posteriors with high correlations. We summarize the algorithm as follows. Suppose the posterior density of $\boldsymbol{\beta}$ is $\pi(\boldsymbol{\beta})$, known up to a proportionality constant. Throughout, we will let $\boldsymbol{\beta}_*$ represent the result of a conventional Markov chain cycle before possible modification.

**Algorithm 1**

Step 1. Starting with $\boldsymbol{\beta}^{(k)}$, generate one full cycle of Gibbs and/or Metropolis-Hastings steps to obtain $\boldsymbol{\beta}_*$;

Step 2. Propose a step of the form $\boldsymbol{\beta}^p = \boldsymbol{\beta}_* + Z\boldsymbol{g}$, where $\boldsymbol{g}$ ranges over a suitable space $\mathcal{S}$ and $Z$ is an independent draw from a convenient symmetric distribution.

Step 3. Using the Metropolis rule, accept $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^p$ with probability $\min\{1, \pi(\boldsymbol{\beta}^p)/\pi(\boldsymbol{\beta}_*)\}$. Otherwise set $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}_*$.

If necessary, Steps 2 and 3 can be repeated in each cycle to sample in multiple directions in $\mathcal{S}$. Some tuning is generally needed to find the most effective distribution for $Z$. The fact that the extra step or steps preserves the posterior distribution $\pi(\boldsymbol{\beta})$ is a consequence of Theorem 2 of Liu & Sabatti (2000) applied to the translation group.

This algorithm is anticipated in the hybrid sampler of Nobile (1998) for improved MCMC in a multinomial probit model. In its simplest form, the likelihood for Nobile's model is invariant with respect to scalar changes in the parameter vector, say $\boldsymbol{\theta}$, of the form $c\boldsymbol{\theta}$ for $c > 0$. After each Gibbs cycle, Nobile added an additional Metropolis-Hastings step with proposal $c\boldsymbol{\theta}$ for $c \sim$ Exp(1) and found that convergence was greatly accelerated. Graves' method exploits location invariance for likelihoods arising from linear models in the parameters instead of scale invariance.

To better understand the behavior of Algorithm 1, we were led to consider a Gibbs alternative proposed by Liu & Sabatti (2000). For simplicity, suppose $\dim(\mathcal{S}) = 1$, and let $\boldsymbol{g}$ denote a spanning vector.

**Algorithm 2**

Step 1. Generate one full Gibbs cycle from $\boldsymbol{\beta}^{(k)}$, denoted by $\boldsymbol{\beta}_*$;

Step 2. Sample $Z$ with density proportional to $\pi(\boldsymbol{\beta}_* + \boldsymbol{g}z)$ and set $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}_* + \boldsymbol{g}Z$.

Again, Step 2 can be repeated in multiple directions in $\mathcal{S}$ if necessary, or a corresponding single multivariate step can be taken. By Theorem 1 of Liu & Sabatti (2000), the chain $\{\boldsymbol{\beta}^{(k)}\}$ retains the stationarity distribution $\pi$.

To study the behavior of Gibbs sampling with extra steps as in Algorithm 2, we formalize the procedure with an equivalent definition in the spirit of the ASSR algorithm of Liu (2003) as follows. Because of confusion in the role of the parameter $\boldsymbol{\beta}$ in the models treated below, it is convenient to replace $\boldsymbol{\beta}$ with a generic parameter $\boldsymbol{\theta}$. Consider an invertible transformation $\boldsymbol{\gamma} = \boldsymbol{G}\boldsymbol{\theta}$, where

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_1' \\ \boldsymbol{G}_2' \end{pmatrix} \text{ and } \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1' \\ \boldsymbol{\gamma}_2' \end{pmatrix}. \tag{4}$$

The supplemental Gibbs step is to draw a new Gibbs sample $\boldsymbol{\gamma}_{1+}$ from the distribution of $\boldsymbol{\gamma}_1 \mid \boldsymbol{\gamma}_2$. The complete algorithm is summarized as follows.

**Algorithm 2′**

Step 1. Generate one full Gibbs cycle $\boldsymbol{\theta}^{(k)}$, denoted by $\boldsymbol{\theta}_*$;

Step 2. Sample $\boldsymbol{\gamma}_{1+}$ from the distribution of $\boldsymbol{\gamma}_1 = \boldsymbol{G}_1'\boldsymbol{\theta}$ given $\boldsymbol{\gamma}_{2*} = \boldsymbol{G}_2'\boldsymbol{\theta}_*$;

Step 3. $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{G}^{-1}(\boldsymbol{\gamma}_{1+}', \boldsymbol{\gamma}_{2*}')'$.

In the following, we will generally assume

$$\boldsymbol{G}_1'\boldsymbol{G}_2 = \mathbf{0}. \tag{5}$$

Consider the special case where $\boldsymbol{G}$ is orthogonal, i.e., $\boldsymbol{G}'\boldsymbol{G} = \boldsymbol{I}$. If $\pi$ denotes the posterior distribution of $\boldsymbol{\theta}$, Step 2 in Algorithm 2′ is a sample from the distribution with density

$$\begin{aligned}[\boldsymbol{\gamma}_1 \mid \boldsymbol{\gamma}_{2*}] &\propto \pi(\boldsymbol{G}_1\boldsymbol{\gamma}_1 + \boldsymbol{G}_2\boldsymbol{\gamma}_{2*}) \\ &= \pi(\boldsymbol{G}_1(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_{1*}) + \boldsymbol{\theta}_*).\end{aligned}$$

The random walk Metropolis step in Algorithm 1 can be seen as an approximation with proposal $\boldsymbol{G}_1\boldsymbol{z} + \boldsymbol{\theta}_*$ in place of the presumably optimal Gibbs distribution. We also now see that Algorithm 2 is equivalent to Algorithm 2′, since the conditional distribution of $z$ in Algorithm 2 is exactly the distribution of $\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_{1*}$ given $\boldsymbol{\gamma}_{2*}$. The issue is how to effectively choose the subspace spanned by the columns of $\boldsymbol{G}_1$ (or equivalently the columns of $\boldsymbol{G}_2$).

The purpose of this paper is to explore the theoretical justification of these extra cycles. Clearly the two algorithms are closely related. In this paper, our main theoretical results pertain to Algorithm 2′. We believe that these results explain not only the effectiveness of Algorithm 2′ but also of the generalized version Algorithm 1.

The rest of the paper is organized as follows. Section 2 contains two extended numerical examples showing the practical effects of adding decorrelation steps in a standard linear model and logistic regression. In Section 3, we show that the Gibbs step of Algorithm 2 is equivalent to an "alternating subspace-spanning resampling" (ASSR) move as proposed by Liu (2003). This allows us to develop some theory for Algorithm 2 in the case of Gibbs sampling with a Gaussian posterior. We apply the theory to several general classes of hierarchical and mixed linear models, and we demonstrate theoretically how added decorrelation steps in the manner of Algorithm 2 can in principle drive the autocovariance in successive cycles of Gibbs sampler MCMC to zero. The results often can be anticipated by direct examination of the likelihood and priors, making implementation straightforward in a variety of settings. Further application and discussion are presented in Sections 4 and 5.

## 2 Decorrelation Steps in Practice

The basic hybrid Algorithm 1 or 2 adds one or more Gibbs or Metropolis-Hastings steps following each cycle of a standard MCMC algorithm. For linear or generalized linear models, these steps often turn out to be in the null space of the design matrix and are not hard to identify by inspection. To illustrate the effectiveness of the method, we consider two examples. The first is an instance of a classical linear mixed model with conjugate normal and inverse-gamma priors. We demonstrate how the high autocorrelation often seen with ordinary Gibbs sampling can be virtually eliminated by a single well-chosen Gibbs step. The example is meant to be illustrative. For linear models with Gaussian errors, it is usually feasible to generate a block sample from the Gaussian portion of the posterior. However, in very large problems, this strategy may not be practical. The second example demonstrates that our techniques are also effective for non-Gaussian models. We present a simulated data set featuring a two-way ANOVA model with interactions and binomial data. We use Metropolis-Hastings sampling with additional Metropolis moves and achieve good mixing.

**Example 3: Mixed Model Balanced Incomplete Block Design** We illustrate the performance of Gibbs sampling with a Gibbs decorrelation step using data from an experiment examining the effects of six fertilizer treatments on potato yield. The data set is given in Example 9.4.1 of Christensen (1996). The dependent variable is potato yield in pounds. The model is

$$y_{ij} = \alpha_i + \theta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \tau_0^{-1}),$$

where $i = 1, \ldots, 6$ indexes the block and $j \in D_i$ is the set of fertilizer treatments in block $i$. There are a total of 30 observations, with five of the six possible treatments in each block. The block effects are random and the treatment effects are fixed. Because we wanted our prior specification to be noninformative, the fixed part of the model must have full rank and we did not include an extra constant term in the model. We took flat priors on the $\theta_j$ and a hierarchical prior on the random effects, $\alpha_i$ independent $\mathrm{N}(0, \tau_1^{-2})$ with hyperprior $[\tau_1] \propto \tau_1^{-3/2}$. Finally, we took the prior $[\tau_0] \propto \tau_0^{-1}$.

This parameterization is centered since $\alpha_i$ is modeled as a random effect. However, without restriction, the model is still poorly specified because changes in parameters of the form $\alpha_{i*} = \alpha_i + Z$, $\theta_{j*} = \theta_j - Z$, $i, j = 1, \ldots, 6$, leave the likelihood unchanged.

Letting $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_6, \theta_1, \ldots, \theta_6)'$, the null space of the corresponding $\boldsymbol{X}'\boldsymbol{X}$ consists of all vectors of the form $c\boldsymbol{g}_1$, where $\boldsymbol{g}_1 = (1, \ldots, 1, -1, \ldots, -1)'$. This suggests using Algorithm 1 to supplement the usual Gibbs sampling cycle with extra moves in the direction $\boldsymbol{\beta}+Z\boldsymbol{g}_1$ for random $Z$ using the Metropolis rule. Alternatively, after a suitable nonsingular transformation $\boldsymbol{\gamma} = \boldsymbol{G}\boldsymbol{\beta}$ with $\boldsymbol{G}' = [\boldsymbol{g}_1, \boldsymbol{G}_2]$, the extra Gibbs step in Algorithm 2 can be used to sample $\gamma_1$. Either extra step has the effect of shifting the Markov chain parallel to the major axis of the posterior distribution and reducing autocorrelation.

Since conjugate distributions exist for all the full conditionals with this model, we used Gibbs sampling on blocks. We updated the random effects, the fixed effects, and the variance components in turn in each cycle. (This design is not complete, so block updates are theoretically more efficient than individual parameter updates.) The left panels of Figure 1 show trace plots of the MCMC run for selected parameters following 1,000 burn-in cycles. The poor mixing predicted by the nonunique parameterization is clearly evident.

Next, we implemented Algorithm 2$'$ with $\gamma_1 = \boldsymbol{g}_1'\boldsymbol{\beta}$. We computed the required full-rank $11 \times 12$ matrix $\boldsymbol{G}_2'$ such that $\boldsymbol{G}_2'\boldsymbol{g}_1 = \boldsymbol{0}$ by computing the eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$. Step 2 of Algorithm 2$'$ is simple since the full conditional is normal. The corresponding trace plots for 1,000 cycles following a 1,000 cycle burn-in run in the right panels of Figure 1 show almost perfect mixing. The plots on the right for $\alpha_1$ and $\theta_1$ display outliers and rather large posterior variances. These plots demonstrate how sampling with the decorrelation step allows much more efficient exploration of the posterior distribution by the Gibbs sampler. These nearly independent draws from the marginal posterior distributions display the true posterior variability of the posteriors under our weakly informative priors. It would take far longer to see this variability with the naive Gibbs sampler.

The naive sampling scheme works well for sampling from the posterior of $\tau_0$. The indeterminacy in the specification of the $\alpha_i$ and $\theta_j$ individually is not present in estimating $\alpha_i+\theta_j$. Consequently, the residual sum of squares needed to sample from the full conditional posterior of $\tau_0$ is stable. However, the situation with other variance components can be different. The plots at the bottom of Figure 1 clearly show the improved mixing with the extra decorrelation step. As a side benefit not shown here, the burn-in period appears much shorter with decorrelation steps than without.

**Example 4: Logistic Two-way ANOVA with Interactions** The next example was implemented using YADAS (Graves (2003*a*,*b*)), with which the methods discussed in this paper are particularly useful, intuitive and easy to apply. YADAS rarely samples from exact conditional distributions, preferring to use exclusively Metropolis(-Hastings) moves. The first "naive" attempt at an analysis updates each scalar parameter individually, using a random-walk step. When this fails to generate algorithms that mix adequately, YADAS features a `MultipleParameterUpdate` that makes it easy to propose Metropolis–Hastings moves to multiple parameters simultaneously. This approach makes use of the generality of the formula for Metropolis–Hastings acceptance probabilities. This paper discusses moves that are samples from conditional distributions of linear transformations of parameters. In the Metropolis context, the analogous moves feature proposals chosen randomly in directions in the null space of $\boldsymbol{X}'\boldsymbol{X}$, accepted or rejected according to the Metropolis rule.

A generalized linear model example where additional Metropolis moves in directions in the null space of $\boldsymbol{X}'\boldsymbol{X}$ greatly facilitate mixing of the MCMC algorithm is a two-way ANOVA model with interactions and binomial data:

$$\text{logit}(p_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

for $1 \leq i \leq 3$ and $1 \leq j \leq 4$. We placed $N(0, 100)$ prior distributions on each of $\mu$, the $\alpha_i$, the $\beta_j$, and the $\gamma_{ij}$. For each $(i, j)$ we constructed synthetic data $y_{ij}$ from the binomial distribution with sample sizes $n_{ij} \equiv 10$ and probability $p_{ij}$. The model is overparameterized: to explain only twelve means the model uses twenty parameters.

The naive algorithm in which we update each of the twenty parameters in turn using random-walk Metropolis steps mixes poorly, and we improve this mixing by adding several Metropolis steps in directions that do not change the likelihood. After a full cycle of individual parameter updates, we obtain $\alpha_{i*}, \beta_{j*}$, and $\gamma_{ij*}$. Next, for each $i$, we generate an innovation $Z_{i\alpha} \sim N(0, s^2_{i\alpha})$, and propose new values of $\alpha_i$ and $\gamma_{ij}$ according to $\alpha_{i+,1} = \alpha_{i*} + Z_{i\alpha}$ and $\gamma_{ij+,1} = \gamma_{ij*} - Z_{i\alpha}$ for all $j$. After the acceptance or rejection, the new value of $\alpha_i$ will be denoted $\alpha_{i*,1}$, which is equal to either $\alpha_{i*}$ or $\alpha_{i+,1}$, and we use similar notation for $\gamma_{ij*,1}$. Analogously, for each $j$ we generate an innovation $Z_{j\beta} \sim N(0, s^2_{j\beta})$, and propose new values of $\beta_j$ and $\gamma_{ij}$ according to $\beta_{j+,2} = \beta_{j*} + Z_{j\beta}$ and $\gamma_{ij+,2} = \gamma_{ij*,1} - Z_{j\beta}$ for all $i$. The new values of these parameters are denoted by $\beta_{j*,2}$ and $\gamma_{ij*,2}$. Finally, we generate normal random variables $W_\mu, W_\alpha, W_\beta,$ and $W_\gamma$ with $W_\mu + W_\alpha + W_\beta + W_\gamma = 0$, and propose $\mu_{+,3} = \mu_* + W_\mu$, $\alpha_{i+,3} = \alpha_{i*,1} + W_\alpha$ for all $i$, $\beta_{j+,3} = \beta_{j*,1} + W_\beta$ for all $j$, and

$\gamma_{ij+,3} = \gamma_{ij*,2} + W_\gamma$ for all $(i, j)$. This adds up to a total of eight additional Metropolis moves. It can be verified that these moves correspond to updating parameters in the null space of the implicit balanced $\boldsymbol{X}'\boldsymbol{X}$ matrix as in the low data variance case of Theorems 2 and 3. See Figure 2 for trace plots of the results under each of the two algorithms. The $\alpha_i$'s, the $\beta_j$'s, and the $\gamma_{ij}$'s all mix efficiently (the Raftery & Lewis (1996) diagnostics agree) whereas $\mu$ still causes some trouble. The right column features the algorithm that includes the additional update steps. This algorithm takes more time, but only by a factor of about 1.5, and it is clear that two iterations of the improved chain are more valuable than three of the unimproved chain. We also stress that moves of this form are easy to implement, especially in YADAS with its `MultipleParameterUpdate` construct.

# 3 Gibbs Decorrelating Steps in Gaussian Linear Models

This section contains the main theoretical results underlying the use of decorrelating steps in linear models. While Algorithms 1 and 2 in principle apply to arbitrary posterior distributions, in this section we focus on Algorithm 2 applied to Gibbs sampling in linear models with Gaussian errors. While these results are interesting in their own right, we believe that they also shed light on the behavior of decorrelation steps in situations when the posterior is only approximately Gaussian such as hierarchical generalized linear models (see Sahu & Roberts 1999).

Subsection 3.1 reviews the basic Gibbs sampling setup of Roberts & Sahu (1997) and contains the key result, Theorem 1, which demonstrates that the draws in the chain $\{\theta^{(k)}\}$ can be made independent under certain circumstances with the proper choice of $\boldsymbol{G}_2$. The remainder of this section is devoted to applications where we show that the intuitive choices for decorrelating steps suggested in the introduction can lead to near perfect decorrelation in a variety of applications with linear and mixed linear models.

## 3.1 Gibbs Decorrelation Steps with a Gaussian Posterior

To exploit the form of the autocorrelation between successive Gibbs cycles, we examine a general case of sampling from a multivariate normal distribution. The setting and notation are adapted from Roberts & Sahu (1997). Suppose $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_r')$ has block structure with (posterior) distribution $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q}$ be the precision matrix, and assume both have block structure commensurate with $\boldsymbol{\theta}$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1r} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{r1} & \boldsymbol{\Sigma}_{r2} & \cdots & \boldsymbol{\Sigma}_{rr} \end{pmatrix} \text{ and } \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} & \cdots & \boldsymbol{Q}_{1r} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} & \cdots & \boldsymbol{Q}_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Q}_{r1} & \boldsymbol{Q}_{r2} & \cdots & \boldsymbol{Q}_{rr} \end{pmatrix}.$$

Consider sequential Gibbs block updates of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r$, the usual sweep order in MCMC labeled DUGS in Roberts & Sahu (1997). Assuming Gibbs samples from the stationary distribution, we will consider a single cycle and label the updates as follows for this cycle. Generate $\boldsymbol{\theta}_{1*}$ given $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r)$, then generate $\boldsymbol{\theta}_{2*}$ given $(\boldsymbol{\theta}_{1*}, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_r)$, etc., with $\boldsymbol{\theta}_* = (\boldsymbol{\theta}_{1*}, \ldots, \boldsymbol{\theta}_{r*})$ denoting the result of one full Gibbs cycle.

Since $\boldsymbol{\theta}_*$ is obtained from $\boldsymbol{\theta}$ from a series of Gibbs steps, each of which has a joint normal distribution with the previous step, it is obvious that the joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_*$ is normal, and the conditional distribution of $\boldsymbol{\theta}_*$ given $\boldsymbol{\theta}$ is also normal. However, this joint distribution is fairly complicated. Roberts & Sahu (1997) gave a concise (if nonintuitive) expression as follows. Following Roberts & Sahu (1997) and Gelfand & Sahu (1999), let $\boldsymbol{Q} = \boldsymbol{L} - \boldsymbol{U}$, where $\boldsymbol{L}$ is the lower block triangular part of $\boldsymbol{Q}$ and $\boldsymbol{U}$ is the negative of the strictly upper block triangular part of $\boldsymbol{Q}$, i.e. $\boldsymbol{L}_{ij} = \mathbf{0}$ for $j > i$ and $\boldsymbol{U}_{ij} = \mathbf{0}$ for $i > j$. Define $\boldsymbol{B} = \boldsymbol{L}^{-1}\boldsymbol{U}$. (Roberts & Sahu (1997) use an equivalent definition of $\boldsymbol{B}$.) Then Roberts & Sahu (1997) showed that the distribution of $\boldsymbol{\theta}_*$ given $\boldsymbol{\theta}$ is $\mathrm{N}(\boldsymbol{B}\boldsymbol{\theta} + (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\mu}, \boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$. Thus the Markov chain is a first order vector-autoregression. The matrix $\boldsymbol{B}$ plays a central role in our work. Note that the condition $E(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}) = \boldsymbol{B}\boldsymbol{\theta} + (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\mu}$ implies

$$\mathrm{Cov}(\boldsymbol{\theta}_*, \boldsymbol{\theta}) = \boldsymbol{B}\boldsymbol{\Sigma}. \tag{6}$$

A useful equivalent way of writing the first order autoregression for updating $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}_* = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{Z}, \tag{7}$$

where $\boldsymbol{Z}$ has a multivariate normal distribution with mean zero and is independent of $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$. Since $\boldsymbol{\theta}_*$ has again the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, $\boldsymbol{Z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$.

By Theorem 1 of Liu & Sabatti (2000), the decorrelation step of Algorithm $2'$ for Gaussian posteriors preserves the Markov nature (7) of the MCMC algorithm. To derive the explicit form of the vector-autoregression, define $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}'$, where $\boldsymbol{G}$ is defined in

(4), and let $\tilde{\boldsymbol{Q}} = \tilde{\boldsymbol{\Sigma}}^{-1}$ be partitioned as

$$\tilde{\boldsymbol{Q}} = \begin{pmatrix} \tilde{\boldsymbol{Q}}_{11} & \tilde{\boldsymbol{Q}}_{12} \\ \tilde{\boldsymbol{Q}}_{21} & \tilde{\boldsymbol{Q}}_{22} \end{pmatrix}. \tag{8}$$

It is not necessary to assume that $\boldsymbol{G}$ is orthogonal. For a general result, let $\boldsymbol{G}^{-1} = \boldsymbol{H} = (\boldsymbol{H}_1, \boldsymbol{H}_2)$ and define

$$\boldsymbol{V} = (\boldsymbol{H}_2 - \boldsymbol{H}_1\tilde{\boldsymbol{Q}}_{11}^{-1}\tilde{\boldsymbol{Q}}_{12})\boldsymbol{G}_2'. \tag{9}$$

With this notation, we have the following extension of Theorem 1 of Roberts & Sahu (1997), whose proof is in the appendix.

**Theorem 1** *Let $\boldsymbol{\theta}_+ = \boldsymbol{G}^{-1}(\boldsymbol{\gamma}_{1+}', \boldsymbol{\gamma}_{2*}')'$, where $\boldsymbol{\gamma}_{1+} \sim (\boldsymbol{\gamma}_1 \mid \boldsymbol{\gamma}_{2*})$ and $\boldsymbol{\gamma}_* = \boldsymbol{G}\boldsymbol{\theta}_*$, with $\boldsymbol{\theta}_*$ given by (7).*

*(a) The correlation between successive updates in one augmented cycle is $Cov(\boldsymbol{\theta}_+, \boldsymbol{\theta}) = \boldsymbol{V}\boldsymbol{B}\boldsymbol{\Sigma}$.*

*(b) The transition law is $(\boldsymbol{\theta}_+ \mid \boldsymbol{\theta}) \sim\ N((\boldsymbol{I} - \boldsymbol{V}\boldsymbol{B})\boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{B}\boldsymbol{\theta}, \boldsymbol{\Sigma} - \boldsymbol{V}\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'\boldsymbol{V}')$.*

The goal of the extra update step is to reduce the autocorrelation between successive updates. With the aid of the theorem, we see that it is theoretically possible to entirely eliminate this autocorrelation by a proper choice of transformation $\boldsymbol{G}$.

**Corollary 1** *If $\boldsymbol{G}_2'\boldsymbol{B} = \boldsymbol{0}$, then*

*(a) $Cov(\boldsymbol{\theta}_+, \boldsymbol{\theta}) = \boldsymbol{0}$, and*

*(b) the transition law is i.i.d. sampling $(\boldsymbol{\theta}_+ \mid \boldsymbol{\theta}) \sim\ N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

The motivation for the decorrelation step is that the matrix $\boldsymbol{B}$ may essentially have low rank. In that case, one would take the columns of $\boldsymbol{G}_1$ to span the most important part of the column space of $\boldsymbol{B}$. If (5) holds, then Algorithm 2$'$ will be highly effective. In some cases, it may be worthwhile to compute $\boldsymbol{B}$ directly and construct a good transformation to reduce $\boldsymbol{G}_2'\boldsymbol{B}$ by examining the singular value decomposition of $\boldsymbol{B}$. In other cases, it is possible to achieve good results simply by inspecting the problem. In the next sections, we explore cases where it is possible to specify good decorrelation steps *a priori*.

### 3.2 Gibbs Sampling in Linear Mixed Models: Low Data Variance

We first restrict attention to the Gaussian linear model with fixed error variance and conjugate priors. Using the framework of Gelfand & Sahu (1999), we show that a decorrelation step or steps in the null space of $\boldsymbol{X}$ leads to asymptotically uncorrelated successive Gibbs samples when the data variance goes to zero.

Consider the standard Gaussian linear model

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{10}$$

with $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \tau_0^{-1}\boldsymbol{I}_n)$. Assume the conjugate prior for $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}) \;\propto\; |\boldsymbol{M}(\boldsymbol{\tau})|_+^{1/2} \exp\Big\{-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{\beta}\Big\}, \tag{11}$$

where $|\boldsymbol{M}|_+$ denotes the product of the positive eigenvalues of $\boldsymbol{M}$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_r)'$ is a vector of (known) precisions (i.e. inverses of variance components) associated with the terms of the linear model ($\tau_i = 0$ is possible). Typically, $\boldsymbol{M}(\boldsymbol{\tau})$ is a block-diagonal matrix,

$$\boldsymbol{M}(\boldsymbol{\tau}) = \mathrm{diag}(\tau_1\boldsymbol{M}_1, \ldots, \tau_r\boldsymbol{M}_r) \tag{12}$$

for nonnegative matrices $\boldsymbol{M}_i$. For example, consider the two-way additive model $y_{ijk} = \mu + \alpha_{1i} + \alpha_{2i} + \varepsilon_{ijk}$, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, $k = 1, \ldots, n_{ij}$. Let $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3]$, where $\boldsymbol{X}_1 = (1, \ldots, 1)'$, and in matrix notation write $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_1\mu + \boldsymbol{X}_2\boldsymbol{\alpha}_1 + \boldsymbol{X}_3\boldsymbol{\alpha}_2$. Thus $\boldsymbol{\beta}' = (\mu, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')$, and the prior is $\mu \sim N(0, \tau_1^{-1})$, $\boldsymbol{\alpha}_1 \sim N(\boldsymbol{0}, \tau_2^{-1}\boldsymbol{I})$, $\boldsymbol{\alpha}_2 \sim N(\boldsymbol{0}, \tau_3^{-1}\boldsymbol{I})$.

Assuming $\tau_0\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}(\boldsymbol{\tau})$ has full rank, the posterior distribution of $\boldsymbol{\beta}$ given $(\tau_0, \boldsymbol{\tau}, \boldsymbol{y})$ is

$$(\boldsymbol{\beta} \mid \tau_0, \boldsymbol{\tau}, \boldsymbol{y}) \sim N(\tilde{\boldsymbol{\beta}}, (\tau_0\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}(\boldsymbol{\tau}))^{-1}), \tag{13}$$

where

$$\tilde{\boldsymbol{\beta}} = (\tau_0\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}(\boldsymbol{\tau}))^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{14}$$

In the general case, let $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r]$, where each $\boldsymbol{X}_i$ is a full rank $n \times m_i$ matrix, and let $m = m_1 + \cdots + m_r$ be the total number of parameters in the linear model. Assume further that $\boldsymbol{M}(\boldsymbol{\tau})$ is block-diagonal as in (12). We examine the case $\tau_0 \to \infty$ (i.e., data variance goes to zero), and consider the matrix $\boldsymbol{B}_{\tau_0}$ for fixed $\boldsymbol{\tau}$ as a function of $\tau_0$. Let

$\boldsymbol{Q}_{\tau_0} = \tau_0\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}(\boldsymbol{\tau}) = \boldsymbol{L}_{\tau_0} - \boldsymbol{U}_{\tau_0}$ as before, where $\boldsymbol{L}_{\tau_0}$ is lower block triangular, and let $\boldsymbol{B}_{\tau_0} = \boldsymbol{L}_{\tau_0}^{-1}\boldsymbol{U}_{\tau_0}$. Finally, define $\boldsymbol{Q} = \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{L} - \boldsymbol{U}$, where again $\boldsymbol{L}$ is lower block triangular. Then $\boldsymbol{B} = \lim_{\tau_0\to\infty}\boldsymbol{B}_{\tau 0} = \boldsymbol{L}^{-1}\boldsymbol{U}$. Suppose $\boldsymbol{G}_2$ is chosen so that the columns of $\boldsymbol{G}_2$ span $\mathcal{C}(\boldsymbol{X}')$. Gelfand & Sahu (1999) considered the lower dimensional parameter $\boldsymbol{\delta} = \boldsymbol{G}_2'\boldsymbol{\beta}$. If

$$\boldsymbol{Q}\boldsymbol{L}^{-1}\boldsymbol{Q} = \boldsymbol{Q}, \tag{15}$$

they showed that the Gibbs sampler on $\boldsymbol{\delta}$ behaves well, and in fact, in the low data variance case as $\tau_0 \to \infty$, successive iterates tend to be independent. We have the following closely related theorem. In the following, the notation $\mathcal{C}(\boldsymbol{A})$ and $\mathcal{N}(\boldsymbol{A})$ refers to the column and null space respectively of a matrix $\boldsymbol{A}$.

**Theorem 2** *Suppose (15) holds. Then*

$$\mathcal{C}(\boldsymbol{B}) \subseteq \mathcal{N}(\boldsymbol{X}'). \tag{16}$$

*Assume further that $\mathcal{C}(\boldsymbol{G}_2) \subseteq \mathcal{C}(\boldsymbol{X}')$ so that $\mathcal{N}(\boldsymbol{X}') \subseteq \mathcal{C}(\boldsymbol{G}_1)$. Then if $\boldsymbol{\gamma}_{1+}$ is sampled from the distribution of $\boldsymbol{\gamma}_1 = \boldsymbol{G}_1'\boldsymbol{\theta}$ given $\boldsymbol{\gamma}_{2*} = \boldsymbol{G}_2'\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}_+ = \boldsymbol{G}^{-1}({\boldsymbol{\gamma}_{1+}}', {\boldsymbol{\gamma}_{2*}}')'$,*

$$\lim_{\tau_0\to\infty} Cov(\boldsymbol{\theta}_+, \boldsymbol{\theta}) = \boldsymbol{0}. \tag{17}$$

**Proof.** By Corollary 1, (17) is true if $\lim_{\tau_0\to\infty}\boldsymbol{G}_2'\boldsymbol{B}_{\tau_0} = \boldsymbol{0}$. Under the assumption of the theorem, it suffices to show $\lim_{\tau_0\to\infty}\boldsymbol{X}\boldsymbol{B}_{\tau_0} = \boldsymbol{0}$. By continuity, $\lim_{\tau_0\to\infty}\boldsymbol{X}\boldsymbol{B}_{\tau_0} = \boldsymbol{X}\boldsymbol{B}$, so it suffices to show (16). But $\boldsymbol{Q}\boldsymbol{L}^{-1}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{L}^{-1}(\boldsymbol{L}-\boldsymbol{U}) = \boldsymbol{Q} - \boldsymbol{Q}\boldsymbol{B}$. Since

$$\boldsymbol{X}\boldsymbol{B} = \boldsymbol{0} \tag{18}$$

if and only if $\boldsymbol{Q}\boldsymbol{B} = \boldsymbol{0}$, (16) holds if and only if $\boldsymbol{L}$ is a generalized inverse of $\boldsymbol{Q}$, i.e. (15) holds.

□

The theorem gives an abstract answer to the question of finding an appropriate extra parameter to define the decorrelation Gibbs step in Algorithm 2′ for the low data variance case with linear models under the generalized inverse condition (15). The extra step should be taken in the null space of $\boldsymbol{X}'\boldsymbol{X}$. To apply this result, one needs a readily available criterion to verify that $\boldsymbol{L}^{-1}$ is a generalized inverse of $\boldsymbol{Q}$. Gelfand & Sahu (1999) claimed that the generalized inverse property holds for any matrix of the form

$$\boldsymbol{X} = (\boldsymbol{X}_0\boldsymbol{\Delta}_1, \boldsymbol{X}_0\boldsymbol{\Delta}_2, \ldots, \boldsymbol{X}_0\boldsymbol{\Delta}_{s-1}, \boldsymbol{X}_0), \tag{19}$$

where $\boldsymbol{X}_0$ has full rank. This class of design matrices includes ANOVA models that are fully nested or main-effects models that include all interactions. We show that certain balanced additive models with or without interactions also share this interesting property. The proof is contained in the appendix.

**Theorem 3** *Let $\boldsymbol{P}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$ denote the perpendicular projection matrix on the column space of $\boldsymbol{X}_i$ for $i = 1, \dots, r$. If*

$$\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{P}_j\boldsymbol{P}_i, \quad 1 \le i, j \le r, \tag{20}$$

*then $\boldsymbol{Q}\boldsymbol{L}^{-1}\boldsymbol{Q} = \boldsymbol{Q}$.*

To summarize, if either condition (19) or condition (20) holds, adding a decorrelation Gibbs step in the null space of $\boldsymbol{X}'\boldsymbol{X}$ will produce perfect sampling in the limit as the data variance goes to zero.

To illustrate how Theorem 2 works, we examine two simple models where we can actually calculate $\boldsymbol{B}$.

**Example 1 (cont.)** Consider first the simplified one-way effects model, $y_i = \mu + \alpha_i + \varepsilon_i$, $i = 1, \dots, m$, with independent components $\varepsilon_i \sim \mathrm{N}(0, \tau_0^{-1})$, $\alpha_i \sim \mathrm{N}(0, \tau_1^{-1})$ and $\mu \sim \mathrm{N}(0, \tau_2^{-1})$. (Note that $\tau_2 = 0$ is possible.). Letting $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{I})$ be the corresponding $m \times (m+1)$ design matrix, the posterior precision matrix is

$$\boldsymbol{Q}(\boldsymbol{\tau}) = \begin{pmatrix} m\tau_0 + \tau_2 & \tau_0\boldsymbol{1}' \\ \tau_0\boldsymbol{1} & (\tau_0 + \tau_1)\boldsymbol{I} \end{pmatrix}.$$

By direct calculation,

$$\lim_{\tau_0 \to \infty} \boldsymbol{B}_{\tau_0} = \lim_{\tau_0 \to \infty} (m + \tau_2/\tau_0)^{-1} \begin{pmatrix} 0 & -\boldsymbol{1}' \\ \boldsymbol{0} & (1 + \tau_1/\tau_0)^{-1}\boldsymbol{1}\boldsymbol{1}' \end{pmatrix} = \begin{pmatrix} 0 & -\boldsymbol{1}' \\ \boldsymbol{0} & \boldsymbol{1}\boldsymbol{1}' \end{pmatrix}.$$

This shows that the limiting autocovariance between successive complete Gibbs cycles is nonzero. However, an ASSR step with $\boldsymbol{G}_2 \perp (-1, 1, \dots, 1)'$ will achieve perfect limiting decorrelation. The natural way to achieve this is to take $\boldsymbol{G}_1 = (-1, 1, \dots, 1)'$ and find $\boldsymbol{G}_2$ such that $\boldsymbol{G}_2'\boldsymbol{G}_1 = \boldsymbol{0}$. In other words, let $\gamma_1 = (-1, 1, \dots, 1)'(\mu, \alpha_1, \dots, \alpha_m) = -\mu + \sum_{i=1}^m \alpha_i$. This corresponds exactly to the extra Metropolis move suggested by direct inspection of the likelihood in the introduction.

**Example 5: Two-way ANOVA** Next, consider the balanced two-way ANOVA effects model

$$y_{ij} = \mu + \beta_{1i} + \beta_{2j} + \varepsilon_{ij}, i = 1, \dots, s; j = 1, \dots, t,$$

with $\mu \sim \text{N}(0, 1/\tau_1)$, $\beta_{1i} \overset{\text{iid}}{\sim} \text{N}(0, 1/\tau_2)$ for $i = 1, \dots, s$, and $\beta_{2j} \overset{\text{iid}}{\sim} \text{N}(0, 1/\tau_3)$ for $j = 1, \dots, t$. Letting $\boldsymbol{\beta}' = (\mu, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, without loss of generality, assume $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2)$, where $\boldsymbol{X}_1 = \boldsymbol{I}_s \otimes \mathbf{1}_t$ and $\boldsymbol{X}_2 = \mathbf{1}_s \otimes \boldsymbol{I}_t$. One can show that the precision matrix of the posterior is

$$\boldsymbol{Q} = \tau_0 \begin{pmatrix} st + \tau_1/\tau_0 & t\mathbf{1}_s' & s\mathbf{1}_t' \\ t\mathbf{1}_s & (t + \tau_2/\tau_0)\boldsymbol{I}_s & \mathbf{1}_s\mathbf{1}_t' \\ s\mathbf{1}_t & \mathbf{1}_t\mathbf{1}_s' & (s + \tau_3/\tau_0)\boldsymbol{I}_t \end{pmatrix}.$$

Consequently, letting $\tau_0 \to \infty$ to model the low data variance case,

$$\lim_{\tau_0 \to \infty} \frac{1}{\tau_0}\boldsymbol{Q} = \begin{pmatrix} st & t\mathbf{1}_s' & s\mathbf{1}_t' \\ t\mathbf{1}_s & t\boldsymbol{I}_s & \mathbf{1}_s\mathbf{1}_t' \\ s\mathbf{1}_t & \mathbf{1}_t\mathbf{1}_s' & s\boldsymbol{I}_t \end{pmatrix},$$

and it is not hard to show that

$$\lim_{\tau_0 \to \infty} \boldsymbol{B}_{\tau_0} = \begin{pmatrix} 0 & -\frac{1}{s}\mathbf{1}_s' & -\frac{1}{t}\mathbf{1}_t' \\ \mathbf{0} & \frac{1}{s}\mathbf{1}_s\mathbf{1}_s' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{t}\mathbf{1}_t\mathbf{1}_t' \end{pmatrix}.$$

Thus the column space of the limiting $\boldsymbol{B}$ matrix is spanned by the two vectors $(-1, \mathbf{1}_s', \mathbf{0}')'$ and $(-1, \mathbf{0}', \mathbf{1}_t')'$ as predicted by the theorem, giving decorrelation step parameters $\gamma_1 = -\mu + \sum_{i=1}^{s} \beta_{1i}$ and $\gamma_2 = -\mu + \sum_{j=1}^{t} \beta_{2j}$. A joint Gibbs update of $(\gamma_1, \gamma_2)$ will reduce autocorrelation in the MCMC chain for the low variance case.

### 3.3 Gibbs Sampling in Linear Mixed Models: High Data Variance

The high data variance limiting case is easier, since Gibbs sampling in the limit is sampling from the prior. The following result pertains to the case where the prior in (11) is proper. Since $\boldsymbol{M}(\boldsymbol{\tau})$ is block diagonal, Gibbs sampling in blocks from the posterior produces perfect, independent samples.

**Theorem 4** *Under model (10) with prior (11) and* $\tau_i > 0$, $i = 1, \dots, r$,

$$\lim_{\tau_0 \to 0} \boldsymbol{B}_{\tau_0} = \mathbf{0}.$$

**Proof.** Under the assumptions of the theorem, $\lim_{\tau_0 \to 0} \boldsymbol{L}_{\tau_0} = \boldsymbol{M}(\boldsymbol{\tau})$, and $\lim_{\tau_0 \to 0} \boldsymbol{U}_{\tau_0} = \boldsymbol{0}$. □

### 3.4 Gibbs Sampling in Hierarchical Gaussian Models

The setup for hierarchical centering of Gelfand et al. (1995) is a special case of the hierarchical model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \tau_0^{-1}\boldsymbol{I}), \\ \boldsymbol{\beta} &= \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{M}(\boldsymbol{\tau})^{-1}), \\ \boldsymbol{\alpha} &\sim \mathrm{N}(\boldsymbol{0}, \tau_{r+1}^{-1}\boldsymbol{M}_{r+1}^{-1}), \end{aligned} \tag{21}$$

where $\boldsymbol{M}(\boldsymbol{\tau})$ is again given by (12). We also assume $\boldsymbol{M}_{r+1}$ is a fixed, known matrix. For Gibbs sampling, perform block updates on the components of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_r')'$. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r)$ as before, and again assume conjugate independent priors

$$\begin{aligned} \boldsymbol{\beta}_k \mid \boldsymbol{\alpha} &\sim \mathrm{N}(\boldsymbol{Z}_k\boldsymbol{\alpha}, \tau_k^{-1}\boldsymbol{M}_k^{-1}), \quad k = 1, \ldots, q \\ \boldsymbol{\beta}_k \mid \boldsymbol{\alpha} &\sim \mathrm{N}(\boldsymbol{0}, \tau_k^{-1}\boldsymbol{M}_k^{-1}), \quad k = q+1, \ldots, r \end{aligned} \tag{22}$$

for some $1 \leq q \leq r$. In this case, $\boldsymbol{Z}$ has a block diagonal form (although not square),

$$\boldsymbol{Z} = \mathrm{diag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q),$$

and $\boldsymbol{M}(\boldsymbol{\tau})$ is given again by (12). Finally, assume $\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z}$ and $\boldsymbol{M}_{r+1}$ have the same block diagonal structure coinciding with the blocks of $\boldsymbol{\alpha}$ being updated. Typically, the $\boldsymbol{\alpha}_k$ are scalars, the $\boldsymbol{Z}_k$ are column vectors, and $\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z}$ and $\boldsymbol{M}_{r+1}$ are diagonal matrices.

In the notation of Section 3.1, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ and $\boldsymbol{\theta}_*$ is the result of one complete cycle of Gibbs sampling applied to the components of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. To study the autocorrelation in the Gibbs samples as a function of data precision $\tau_0$, fix the other variance components $\tau_1, \ldots, \tau_r, \tau_{r+1}$. Since $\mathrm{Cov}(\boldsymbol{\theta}_*, \boldsymbol{\theta}) = \boldsymbol{B}_{\tau_0}\boldsymbol{\Sigma}_{\tau_0}$ again from (6), we examine the behavior of $\boldsymbol{B}_{\tau_0}$ as a function of $\tau_0$.

One can see that the precision matrix of the posterior is

$$\boldsymbol{Q}_{\tau_0} = \begin{pmatrix} \tau_0\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}(\boldsymbol{\tau}) & -\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} \\ -\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau}) & \boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} + \tau_{r+1}\boldsymbol{M}_{r+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}_{xx,\tau_0} & \boldsymbol{Q}_{xz} \\ \boldsymbol{Q}_{zx} & \boldsymbol{Q}_{zz}, \end{pmatrix}, \tag{23}$$

where $\boldsymbol{Q}_{xx,\tau_0} = \tau_0\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}(\boldsymbol{\tau})$, etc. (Note that only the upper left block depends on $\tau_0$).

**Theorem 5** *In model (21) and (22), suppose $\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z}$ and $\boldsymbol{M}_{r+1}$ are block diagonal. Let $\boldsymbol{Q}_{xx,\tau_0} = \boldsymbol{L}_{xx,\tau_0} - \boldsymbol{U}_{xx,\tau_0}$, where $\boldsymbol{L}_{xx,\tau_0}$ is lower block triangular and $\boldsymbol{U}_{xx,\tau_0}$ is strictly upper block triangular, and define $\boldsymbol{B}_{xx,\tau_0} = \boldsymbol{L}^{-1}_{xx,\tau_0}\boldsymbol{U}_{xx,\tau_0}$. Then the following results hold.*

*(a) For all cases,*

$$\boldsymbol{B}_{\tau_0} = \begin{pmatrix} \boldsymbol{B}_{xx,\tau_0} & -\boldsymbol{L}^{-1}_{xx,\tau_0}\boldsymbol{Q}_{xz} \\ -\boldsymbol{Q}^{-1}_{zz}\boldsymbol{Q}_{zx}\boldsymbol{B}_{xx,\tau_0} & \boldsymbol{Q}^{-1}_{zz}\boldsymbol{Q}_{zx}\boldsymbol{L}^{-1}_{xx,\tau_0}\boldsymbol{Q}_{xz} \end{pmatrix}. \tag{24}$$

*(b) To model the low data variance case, let $\lim_{\tau_0\to\infty} \boldsymbol{B}_{xx,\tau_0} = \boldsymbol{B}_{xx,\infty}$. Then*

$$\lim_{\tau_0\to\infty} \boldsymbol{B}_{\tau_0} = \begin{pmatrix} \boldsymbol{B}_{xx,\infty} & \boldsymbol{0} \\ (\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} + \tau_{r+1}\boldsymbol{M}_{r+1})^{-1}\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{B}_{xx,\infty} & \boldsymbol{0} \end{pmatrix}. \tag{25}$$

*(c) Finally, for the high data variance case,*

$$\lim_{\tau_0\to 0} \boldsymbol{B}_{\tau_0} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{Z} \\ \boldsymbol{0} & (\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} + \tau_{r+1}\boldsymbol{M}_{r+1})^{-1}\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} \end{pmatrix}. \tag{26}$$

The proof is again in the Appendix.

We now consider the two limiting cases in turn.

### 3.4.1 Hierarchical Models: Low Data Variance

Since $\boldsymbol{Q}_{xx,\tau_0}$ is the posterior precision matrix $\boldsymbol{Q}$ for the standard Gaussian linear model (10) and (11), $\boldsymbol{B}_{xx,\tau_0}$ is exactly the autocovariance matrix factor from Section 3. In particular, in the low data variance case, the results of Section 3.2 apply. For the rest of this section, assume that $\boldsymbol{X}$ satisfies condition (15).

To illustrate how Theorem 5(b) can be applied, suppose vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ span $\mathcal{S} = \mathcal{N}(\boldsymbol{X}'\boldsymbol{X})$. By Theorem 2, these vectors also span the range space of $\boldsymbol{B}_{xx,\infty}$. Define $\boldsymbol{v}_j = (\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} + \tau_{r+1})^{-1}\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{u}_j$, and let $\boldsymbol{g}_j = (\boldsymbol{u}_j', \boldsymbol{v}_j')'$, $j = 1, \ldots, k$. If $\boldsymbol{G}_1 = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k]$ and $\boldsymbol{G}_2$ is chosen accordingly, equation (25) implies $\lim_{\tau_0\to\infty} \mathcal{C}(\boldsymbol{B}_{\tau_0}) \subset \mathcal{C}(\boldsymbol{G}_1)$. Thus by Corollary 1 again, a joint supplemental Gibbs update of the new parameters $\gamma_j = \boldsymbol{u}_j'\boldsymbol{\beta} + \boldsymbol{v}_j'\boldsymbol{\alpha}$, $j = 1, \ldots, k$, gives a complete decorrelation step for the limiting low data variance case. A Metropolis-Hastings step or steps in these variables in practice also serves as a useful decorrelation step.

The case when $\boldsymbol{X}$ has full rank is especially easy.

**Corollary 2** *In the hierarchical model (21) and (22), if $\boldsymbol{X}$ has full rank and satisfies (15),* $\lim_{\tau_0 \to \infty} \boldsymbol{B}_{\tau_0} = \boldsymbol{0}$.

**Proof.** Equation (16) implies

$$\boldsymbol{X}'\boldsymbol{B}_{xx,\infty} = \boldsymbol{0}, \tag{27}$$

which means $\boldsymbol{B}_{xx,\infty} = \boldsymbol{0}$ since $\boldsymbol{X}$ has full rank. The corollary follows immediately from (25).

□

**Example 2 (cont.)** This corollary clearly covers the the hierarchical version of one-way ANOVA in (2). Since $\boldsymbol{X}$ has full rank, successive Gibbs full cycles produces independent samples from the posterior in the limit, as shown in Gelfand et al. (1995). It is instructive to obtain the result by direct computation of $\boldsymbol{B}$. Take $r = 1$, $\boldsymbol{\beta}' = (\mu_1, \ldots, \mu_m)$, $\alpha = \mu$, $\boldsymbol{Z} = \boldsymbol{1}_m$, $\boldsymbol{M}(\boldsymbol{\tau}) = \tau_1 \boldsymbol{I}_m$, and $\boldsymbol{M}_2 = 1$. Without loss of generality, let $\tau_0 = n/\delta_0$ and $\boldsymbol{X} = \boldsymbol{I}$. The posterior precision matrix is

$$\boldsymbol{Q} = \begin{pmatrix} (\tau_0 + \tau_1)\boldsymbol{I}_m & -\tau_1 \boldsymbol{1}_m \\ -\tau_1 \boldsymbol{1}_m' & m\tau_1 + \tau_2 \end{pmatrix}.$$

If a block Gibbs update is performed on $\boldsymbol{\beta}$ followed by a Gibbs update of $\alpha$,

$$\boldsymbol{B}_{\tau_0} = \begin{pmatrix} \boldsymbol{0} & \frac{\tau_1}{\tau_0 + \tau_1} \boldsymbol{1}_m \\ \boldsymbol{0} & \frac{\tau_1^2}{(m\tau_1 + \tau_2)(\tau_0 + \tau_1)} \end{pmatrix}.$$

Clearly, as $\tau_0 \to \infty$, $\boldsymbol{B}_{\tau_0} \to \boldsymbol{0}$, verifying that hierarchical centering eradicates the autocorrelation in MCMC cycles. The same conclusion holds in one-way ANOVA if the parameters $\mu_1, \ldots, \mu_r$ are updated individually by Gibbs sampling.

However, in typical models with crossed effects, the null space of $\boldsymbol{X}'\boldsymbol{X}$ is nonempty even without the constant term. Thus hierarchical centering cannot completely eliminate autocorrelation in sucessive Gibbs cycles without further adjustments. Sweeping (Vines et al. 1996) the constant terms from all but one centered effect in an additive model does produce a full-rank $\boldsymbol{X}$ matrix (at the cost of complication in the structure of the prior). Sweeping crossed-effect terms is more complicated. The introduction of suitable decorrelating steps is an easy alternative that retains the structure of simple additive models.

**Example 6: Mixed Model Two-way ANOVA** To illustrate the situation with higher order ANOVA, consider the following balanced complete two-way example. This setup is closely related to the incomplete balanced design of Example 3. Assume $y_{ij} = \beta_{1i} + \beta_{2j} + \varepsilon_{ij}$, $i = 1, \ldots, s; j = 1, \ldots, t$. We have written this model without a constant term, so suppose the prior for the first effect is centered with $\beta_{1i} \overset{\text{iid}}{\sim} \mathrm{N}(\mu, 1/\tau_1)$ for $i = 1, \ldots, s$, and the second effect has prior $\beta_{2j} \overset{\text{iid}}{\sim} \mathrm{N}(0, 1/\tau_2)$ for $j = 1, \ldots, t$. Further assume the prior $\mu \sim \mathrm{N}(0, 1/\tau_3)$. This is an example of a mixed model of the form of Section 3.4 with $q = 1$ and $r = 2$.

Consideration of the likelihood alone as in Example 3 suggests invariance with respect to transformations of the form $\beta_{1i} + Z$, $\beta_{2j} - Z$, $\alpha + Z$ for all $i$ and $j$. Thus Metropolis steps or equivalent Gibbs samples corresponding to the new parameter $\tilde{\gamma} = \boldsymbol{u}_1'\boldsymbol{\beta}$ with $\boldsymbol{u}_1 = (1, \ldots, 1, -1, \ldots, -1, 1)'$ and $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1s}, \beta_{21}, \ldots, \beta_{2t})'$ are plausible. Somewhat surprisingly, the theorem adds a term with the random effect $\alpha$ to this parameter. Specifically, with $\boldsymbol{Z} = (\mathbf{1}_s', \mathbf{0}')$ and $\boldsymbol{M} =$ diag$(\tau_1\boldsymbol{I}_s, \tau_2\boldsymbol{I}_t)$, in this case $v_1 = (\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} + \tau_{r+1})^{-1}\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{u}_1 = s\tau_1/(s\tau_1 + \tau_3)$, so the theoretically optimal decorrelating parameter is in the direction $\boldsymbol{g}_1 = (\boldsymbol{u}_1', v_1)'$, i.e., $\gamma = \sum_{i=1}^{s} \beta_{1i} - \sum_{j=1}^{t} \beta_{2j} + v_1\alpha$. We have found that the decorrelating step using $\tilde{\gamma}$ rather than the optimal $\gamma$ works well in practice.

We can verify this formally by computing the autocovariance factor directly. Without loss of generality, let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ with $\boldsymbol{X}_1 = \boldsymbol{I}_s \otimes \mathbf{1}_t$ and $\boldsymbol{X}_2 = \mathbf{1}_s \otimes \boldsymbol{I}_t$. It follows that the precision matrix of the posterior is explicitly

$$\boldsymbol{Q} = \begin{pmatrix} \tau_0\boldsymbol{X}_1'\boldsymbol{X}_1 + \tau_1\boldsymbol{I}_s & \tau_0\boldsymbol{X}_1'\boldsymbol{X}_2 & -\tau_1\mathbf{1}_s \\ \tau_0\boldsymbol{X}_2'\boldsymbol{X}_1 & \tau_0\boldsymbol{X}_2'\boldsymbol{X}_2 + \boldsymbol{\tau}_2\boldsymbol{I}_t & \mathbf{0} \\ -\tau_1\mathbf{1}_s' & \mathbf{0} & s\tau_1 + \tau_3 \end{pmatrix} = \begin{pmatrix} (t\tau_0 + \tau_1)\boldsymbol{I}_s & \tau_0\mathbf{1}_s\mathbf{1}_t' & -\tau_1\mathbf{1}_s \\ \tau_0\mathbf{1}_t\mathbf{1}_s' & (s\tau_0 + \tau_2)\boldsymbol{I}_t & \mathbf{0} \\ -\tau_1\mathbf{1}_s' & \mathbf{0} & s\tau_1 + \tau_3 \end{pmatrix}.$$

Routine calculations give

$$\boldsymbol{B}_{\tau_0} = \frac{\tau_0}{t\tau_0 + \tau_1} \begin{pmatrix} \mathbf{0} & -\mathbf{1}_s\mathbf{1}_t' & \frac{\tau_1}{\tau_0}\mathbf{1}_s \\ \mathbf{0} & \frac{s\tau_0}{s\tau_0+\tau_2}\mathbf{1}_t\mathbf{1}_t' & -\frac{s\tau_1}{s\tau_0+\tau_2}\mathbf{1}_t \\ \mathbf{0} & -\frac{s\tau_1}{s\tau_1+\tau_3}\mathbf{1}_t' & \frac{s\tau_1^2}{\tau_0(s\tau_1+\tau_3)} \end{pmatrix}.$$

In the limiting case with small data variance ($\tau_0 \to \infty$), take the noninformative prior on $\mu$ with $\tau_3 = 0$. This leads to the limiting autocovariance factor

$$\lim_{\tau_0 \to \infty} \boldsymbol{B}_{\tau_0} = \frac{1}{t} \begin{pmatrix} \mathbf{0} & -\mathbf{1}_s\mathbf{1}_t' & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_t\mathbf{1}_t' & \mathbf{0} \\ \mathbf{0} & -\frac{s\tau_1}{s\tau_1+\tau_3}\mathbf{1}_t' & 0 \end{pmatrix}.$$

Thus even in the limiting low data variance case, $\boldsymbol{B}$ does not vanish, and a further decorrelation step would be useful with new parameter $\gamma = -\sum_{i=1}^{t} \beta_{1i} + \sum_{j=1}^{s} \beta_{2j} - s\tau_1\alpha_1/(s\tau_1+\tau_3)$.

### 3.4.2 Hierarchical Models: High Data Variance

In the usual case with nonsingular $\boldsymbol{Z}$, approximate decorrelating parameters to update for the high data variance case can be obtained directly from Theorem 5(c) by taking

$$\boldsymbol{G}_1 = \begin{pmatrix} \boldsymbol{Z} \\ (\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} + \tau_{r+1}\boldsymbol{M}_{r+1})^{-1}\boldsymbol{Z}'\boldsymbol{M}(\boldsymbol{\tau})\boldsymbol{Z} \end{pmatrix}.$$

Consider Example 2 again with the parameterization from Section 3.4.1. Then $\boldsymbol{G}_1 = (\mathbf{1}_m', m\tau_1/(m\tau_1+\tau_2))'$. A supplemental Metropolis step with proposal $(\beta_1 + Z, \ldots, \beta_m + Z, m\tau_1\alpha/(m\tau_1+\tau_2) + Z)$ or the related Gibbs update of the new parameter $\gamma = \sum_{i=1}^{m} \beta_i + m\tau_1\alpha/(m\tau_1+\tau_2)$ would be effective for improving mixing.

One can anticipate this result when the prior on $\alpha$ is diffuse relative to the prior distribution of the $\beta_i$s (i.e., $\tau_2 \approx 0$) by direct consideration of the unnormalized posterior, which in this case is proportional to

$$\exp\left\{-\frac{\tau_0}{2}\sum_{i=1}^{m}(y_i-\beta_i)^2\right\}\exp\left\{-\frac{\tau_1}{2}\sum_{i=1}^{m}(\beta_i-\alpha)^2\right\}\exp\left\{-\frac{\tau_2\alpha^2}{2}\right\}. \tag{28}$$

As $\tau_0 \to 0$ and $\tau_2 \to 0$, the middle term dominates (28). Thus a Metropolis move with proposal of the form $(\beta_1^C, \ldots, \beta_m^C, \alpha^C) = (\beta_1 + Z, \ldots, \beta_m + Z, \alpha + Z)$, which corresponds to a Gibbs update of the parameter $\gamma_1 = \beta_1 + \cdots + \beta_m + \alpha$, is an appropriate decorrelation step for high data variance.

### 3.4.3 Partially Noncentered Prior

For the special case of one-way ANOVA, one can compute the exact autocovariance factor for Gibbs sampling using prior (3) given by Papaspiliopoulos et al. (2003). With the notation there, $\boldsymbol{\beta} = (\beta_m^w, \ldots, \beta_m^w)'$, $\boldsymbol{X} = \boldsymbol{I}_m \otimes \mathbf{1}_n$ and $\alpha = \mu$, and $\tau_i = 1/\delta_i$, $i = 1, 2, 3$. Then

$$\boldsymbol{Q} = \begin{pmatrix} (n\tau_0+\tau_2)\boldsymbol{I}_m & (n\tau_0 w - \tau_2(1-w))\mathbf{1}_m \\ (n\tau_0 w - \tau_2(1-w))\mathbf{1}_m' & [n\tau_0 w^2 - \tau_2(1-w)]m + \tau_1 \end{pmatrix}$$

and

$$\boldsymbol{B} = \begin{pmatrix} \mathbf{0} & -\frac{n\tau_0 w - \tau_2(1-w)}{tau_0+\tau_2)}\mathbf{1}_m \\ 0 & \frac{m[n\tau_0 w - \tau_2(1-w)]^2}{(n\tau_0+\tau_2)\{[n\tau_0 w^2 - \tau_2(1-w)]m+\tau_1\}} \end{pmatrix}.$$

Taking $w = \tau_2/(n\tau_0 + \tau_2) = \delta_0/(n\delta_2 + \delta_0)$ eradicates the autocovariance between successive Gibbs cycles.

## 4 Full Bayesian Models

In practice, some of the precisions or variances also have prior distributions. Considering the general model (10), we assume conjugate (possibly improper) priors for the $\tau_i$,

$$[\tau_i] \propto \tau_i^{a_i - 1} e^{-b_i \tau_i}, \quad i = 0, \ldots, r,$$

for some real-valued hyperparameters of $(a_i, b_i)$. When $a_i$ and $b_i$ are both positive, the prior is proper. Of course, one or more precision parameters $\tau_i$ could be fixed constants, for example, for "fixed" effects. Since we are mainly concerned with updating the vectors $\boldsymbol{\beta}$ for fixed $\tau_i$, we assume that all $\tau_i$ are random for simplicity. Then the joint density for the likelihood and priors satisfies

$$[\boldsymbol{y} \mid \boldsymbol{\beta}, \tau_0]\,[\boldsymbol{\beta} \mid \boldsymbol{\tau}]\,[\tau_0]\,[\boldsymbol{\tau}] \quad \propto \quad \tau_0^{n/2} \exp\left\{-\frac{\tau_0}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 - \frac{1}{2}\boldsymbol{\beta}' A(\boldsymbol{\tau})\boldsymbol{\beta}\right\} \prod_{i=0}^{r} \tau_i^{a_i - 1} e^{-b_i \tau_i}.$$

By conjugacy, we have the full conditional distributions for $(\tau_0, \boldsymbol{\tau})$,

$$\begin{aligned}(\tau_0 \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}) \quad &\sim \quad \text{Gamma}\Big(a_0 + \frac{n}{2},\, b_0 + \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^{r} \boldsymbol{X}_j \boldsymbol{\beta}_j\Big\|^2\Big), \\ (\tau_i \mid \boldsymbol{y}, \boldsymbol{\beta}, \tau_0, \boldsymbol{\tau}_{-i}) \quad &\sim \quad \text{Gamma}\Big(a_i + \frac{1}{2}m_i,\, b_i + \frac{1}{2}\boldsymbol{\beta}_i' \boldsymbol{M}_i \boldsymbol{\beta}_i\Big), \quad i > 0.\end{aligned}$$

In this case, we can update $\beta$ for fixed $(\tau_0, \boldsymbol{\tau})$ with Gibbs or Metropolis Hastings decorrelation steps. Example 3 of Section 2 illustrates the effectiveness of decorrelation steps in linear models with a full Bayesian analysis when $(\tau_0, \boldsymbol{\tau})$ is treated as random.

## 5 Discussion

Convergence to the stationary distribution and mixing in MCMC algorithms can often be greatly improved by the addition of a one or more relatively simple decorrelation steps to standard Gibbs or Metropolis-Hastings cycles. In this paper, we have shown how these extra moves can facilitate Bayesian computation for distributions involving linear modeling in the parameters. Much of the previous research in the area including Gelfand et al. (1995), Vines et al. (1996), Gelfand et al. (1996) and Papaspiliopoulos et al. (2003) has advocated modified priors and reparameterizations to deal with problems caused by high a posteriori correlation among parameters. In our view, a statistician ought to be able to estimate the posterior resulting from any prior he or she chooses. We believe the methods presented here are a step in that direction.

In many cases, suitable decorrelation steps can be deduced by direct inspection of the likelihood. This observation has led us to consider analysis of many of the traditional linear parameterizations that have less than full rank. Decorrelation steps in the null space of the linear model have proven effective in a number of applied problems. In this paper, we have presented a theoretical analysis with Gaussian distributions and priors coupled with a Gibbs version of the decorrelation step. In doing so, we adapt the Alternating Subspace Spanning Resampling algorithm of Liu (2003) to linear models. The somewhat surprising result is that in the low-data variance case (i.e., with increasingly large sample sizes), MCMC with decorrelation steps approaches "perfect sampling" (Kendall 1998) in that successive Markov chain cycles become asymptotically independent.

Many strategies have been employed to estimate a posteriori quantities by sampling. Different distributions often require different methods, and no single technique has been found to work universally well. Decorrelation steps offer a useful addition to the growing body of methods.

## References

Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995), 'Reply to comments on "Bayesian computation and stochastic systems"', *Statistical Science* **10**, 58–66.

Chib, S. & Carlin, B. P. (1999), 'On MCMC sampling in hierarchical longitudinal models', *Statistics and Computing* **9**(1), 17–26.

Christensen, R. (1996), *Plane answers to complex questions: the theory of linear models*, Springer-Verlag Inc.

Gelfand, A. E. & Sahu, S. K. (1999), 'Identifiability, improper priors, and Gibbs sampling for generalized linear models', *Journal of the American Statistical Association* **94**, 247–253.

Gelfand, A. E., Sahu, S. K. & Carlin, B. P. (1995), 'Efficient parametrisations for normal linear mixed models', *Biometrika* **82**, 479–488.

Gelfand, A. E., Sahu, S. K. & Carlin, B. P. (1996), Efficient parameterizations for generalized linear mixed models, *in* 'Bayesian Statistics 5 – Proceedings of the Fifth Valencia International Meeting', pp. 165–180.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. (2004), 'Parameterization in Bayesian modeling', *Journal of the American Statistical Association* **99**, 537–545.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences (Disc: p483-501, 503-511)', *Statistical Science* **7**, 457–472.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), *Markov chain Monte Carlo in practice*, Chapman & Hall Ltd.

Gilks, W. R. & Roberts, G. O. (1996), Strategies for improving mcmc, *in* 'Markov chain Monte carlo in practice, eds. by W. R. Gilks, S. Richardson and D. J. Spiegelhalter', London: Chapman and Hall, pp. 89–114.

Graves, T. L. (2003*a*), 'A framework for expressing and estimating arbitrary statistical models using Markov chain Monte Carlo', *submitted to Journal of Computational and Graphical Statistics* .

Graves, T. L. (2003*b*), 'An introduction to YADAS', *yadas.lanl.gov* .

Graves, T. L. & Picard, R. R. (2003), 'Seasonal evolution of influenza-related mortality', *Los Alamos National Laboratory Technical Report* **LA-UR-03-1237**.

Graves, T. L., Reese, C. S. & Fitzgerald, M. (2003), 'Hierarchical models for permutations: Analysis of auto racing results', *Journal of the American Statistical Association* **98**, 282–291.

Kendall, W. (1998), Perfect simulation for the area-interaction point process, *in* 'Probability Towards 2000, eds. by Heyde, C. and Accardi, L.', Springer-Verlag, New York, pp. 218–234.

Liu, C. (2003), ‘Alternating subspace-spanning resampling to accelerate Markov chain Monte Carlo simulation’, *Journal of the American Statistical Association* **98**, 110–117.

Liu, J. S. & Sabatti, C. (1999), Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid & A. Smith, eds, ‘Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting’, Clarendon Press [Oxford University Press], pp. 389–413.

Liu, J. S., Wong, W. H. & Kong, A. (1994), ‘Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes’, *Biometrika* **81**, 27–40.

Liu, J. S. & Wu, Y. N. (1999), ‘Parameter expansion for data augmentation’, *Journal of the American Statistical Association* **94**, 1264–1274.

Liu, J. & Sabatti, C. (2000), ‘Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation’, *Biometrika* **87**(2), 353–369.

Meng, X.-L. & van Dyk, D. A. (1999), ‘Seeking efficient data augmentation schemes via conditional and marginal augmentation’, *Biometrika* **86**, 301–320.

Nobile, A. (1998), ‘A hybrid Markov chain for the Bayesian analysis of the multinomial probit model’, *Statistics and Computing* **8**, 229–242.

Papaspiliopoulos, O. (2003), Non-centred parameterisations for hierarchical models and data augmentation. PhD thesis, Department of Mathematics and Statistics, Lancaster University, Lancaster.

Papaspiliopoulos, O., Roberts, G. O. & Sköld, M. (2003), Non-centered parameterizations for hierarchical models and data augmentation (with discussions), *in* ‘Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting, eds. by J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckman, A.F.M. Smith, and M. West’, Clarendon Press [Oxford University Press], pp. 307–326.

Raftery, A. E. & Lewis, S. M. (1996), Implementing MCMC, *in* ‘*Markov chain Monte Carlo in practice*, Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds)’, Chapman & Hall, 115-130.

Roberts, G. O. & Sahu, S. K. (1997), ‘Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler’, *Journal of the Royal Statistical Society, Series B, Methodological* **59**, 291–317.

Sahu, S. K. & Roberts, G. O. (1999), ‘On convergence of the EM algorithm and the Gibbs sampler’, *Statistics and Computing* **9**(1), 55–64.

Tierney, L. (1994), ‘Markov chains for exploring posterior distributions (Disc: p1728-1762)’, *The Annals of Statistics* **22**, 1701–1728.

Vines, S. K., Gilks, W. R. & Wild, P. (1996), ‘Fitting Bayesian multiple random effects models’, *Statistics and Computing* **6**, 337–346.

## 6 Appendix

**Proof of Theorem 1** The decorrelation step samples $\boldsymbol{\gamma}_{1+}$ from the distribution $\boldsymbol{\gamma}_1 \mid \boldsymbol{\gamma}_{2*}$. Let

$$\tilde{\boldsymbol{\mu}} = \begin{pmatrix} \tilde{\boldsymbol{\mu}}_1 \\ \tilde{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}_1'\boldsymbol{\mu} \\ \boldsymbol{G}_2'\boldsymbol{\mu} \end{pmatrix}.$$

Then $\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_{2*} \sim \mathrm{N}(\tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{Q}}_{11}^{-1}\tilde{\boldsymbol{Q}}_{12}(\boldsymbol{\gamma}_{2*} - \tilde{\boldsymbol{\mu}}_2), \tilde{\boldsymbol{Q}}_{11}^{-1})$, or equivalently,

$$\boldsymbol{\gamma}_{1+} \;=\; \tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{Q}}_{11}^{-1}\tilde{\boldsymbol{Q}}_{12}(\boldsymbol{\gamma}_{2*} - \tilde{\boldsymbol{\mu}}_2) + \boldsymbol{Z}_1^+,$$

where $\boldsymbol{Z}_1^+ \sim \mathrm{N}(\boldsymbol{0}, \tilde{\boldsymbol{Q}}_{11}^{-1})$. Using $\boldsymbol{G}^{-1} = \boldsymbol{H} = (\boldsymbol{H}_1, \boldsymbol{H}_2)$,

$$\begin{aligned} \boldsymbol{\theta}_+ &= \boldsymbol{H}_1\boldsymbol{\gamma}_{1+} + \boldsymbol{H}_2\boldsymbol{\gamma}_{2*} \\ &= \boldsymbol{H}_1\{\boldsymbol{G}_1'\boldsymbol{\mu} - \tilde{\boldsymbol{Q}}_{11}^{-1}\tilde{\boldsymbol{Q}}_{12}\boldsymbol{G}_2'(\boldsymbol{\theta}_* - \boldsymbol{\mu}) + \boldsymbol{Z}_1^+\} + \boldsymbol{H}_2\boldsymbol{G}_2'\boldsymbol{\theta}_*. \end{aligned}$$

Noting that $\boldsymbol{H}_1\boldsymbol{G}_1' + \boldsymbol{H}_2\boldsymbol{G}_2' = \boldsymbol{I}$,

$$\boldsymbol{\theta}_+ = (\boldsymbol{I} - \boldsymbol{V})\boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{\theta}_* + \boldsymbol{H}_1\boldsymbol{Z}_{1+},$$

where $\boldsymbol{V}$ is given in (9). Now substitute (7) and simplify to obtain

$$\boldsymbol{\theta}_+ = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{B})\boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{Z}_+,$$

where $\boldsymbol{Z}^+ = \boldsymbol{V}\boldsymbol{Z} + \boldsymbol{H}_1\boldsymbol{Z}_1^+$ is independent of $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}_+$ has the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, Theorem 1 follows.

To prove Theorem 3, we first need the following simple lemma.

**Lemma 1** *If (20) holds and $P_{1:j}$ denotes the perpendicular projection matrix onto the column span of $X_1, \ldots, X_j$ for $j = 1, \ldots, r$, then*

$$P_i P_{1:j} = P_{1:j} P_i \tag{29}$$

*for $i = 1, \ldots, r$.*

**Proof.** The proof is by induction. By condition (20), the result is true for $j = 1$. Suppose (29) is true for some $j = k$, $1 \leq k < r$. Using the induction hypothesis, it is easy to check that

$$\begin{aligned} \boldsymbol{P}_{1:(k+1)} &= \boldsymbol{P}_{1:k} + \boldsymbol{P}_{k+1} - \boldsymbol{P}_{1:k}\boldsymbol{P}_{k+1} \\ &= \boldsymbol{P}_{1:k} + \boldsymbol{P}_{k+1} - \boldsymbol{P}_{k+1}\boldsymbol{P}_{1:k}, \end{aligned} \tag{30}$$

since the right side is symmetric and idempotent with range containing $col(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k+1})$. But by the induction hypothesis, the right side of (30) commutes with each $\boldsymbol{P}_i$, so (29) holds for $j = k + 1$. □

**Proof of Theorem 3.** We prove (18) or equivalently (15) by induction. Define $\boldsymbol{Q}_i = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i)'(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i) = \boldsymbol{L}_i - \boldsymbol{U}_i$ for $i = 1, \ldots, r$, and let $\boldsymbol{X}_{-i} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1})$.

If $r = 1$, then $\boldsymbol{L} = \boldsymbol{Q}$ and (15) is obvious. Suppose now that (15) holds for given model $\boldsymbol{X} = \boldsymbol{X}_{-i} = (\boldsymbol{X}_1 \cdots \boldsymbol{X}_{i-1})$, i.e., $\boldsymbol{Q}_{i-1}\boldsymbol{L}_{i-1}^{-1}\boldsymbol{Q}_{i-1} = \boldsymbol{Q}_{i-1}$. By construction,

$$\boldsymbol{L}_i = \begin{pmatrix} \boldsymbol{L}_{i-1} & \boldsymbol{0} \\ \boldsymbol{X}_i'\boldsymbol{X}_{-i} & \boldsymbol{X}_i'\boldsymbol{X}_i \end{pmatrix} \quad \text{and} \quad \boldsymbol{U}_i = \begin{pmatrix} \boldsymbol{U}_{i-1} & \boldsymbol{X}_{-i}'\boldsymbol{X}_i \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Now (15) holds for $\boldsymbol{L}_i$ if and only if $\boldsymbol{Q}_i\boldsymbol{L}_i^{-1}\boldsymbol{U}_i = \boldsymbol{0}$, or equivalently if and only if $(\boldsymbol{X}_{-i}, \boldsymbol{X}_i)\boldsymbol{L}_i^{-1}\boldsymbol{U}_i = \boldsymbol{0}$. Using the formula for an invertible $2 \times 2$ lower triangular block diagonal matrix

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} & \boldsymbol{0} \\ -\boldsymbol{C}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} & \boldsymbol{C}^{-1} \end{pmatrix}, \tag{31}$$

it is not hard to show that

$$(\boldsymbol{X}_{-i}, \boldsymbol{X}_i)\boldsymbol{L}_i^{-1}\boldsymbol{U}_i = \left((\boldsymbol{I} - \boldsymbol{P}_r)\boldsymbol{X}_{-i}\boldsymbol{L}_{i-1}^{-1}\boldsymbol{U}_{i-1}, \quad -(\boldsymbol{I} - \boldsymbol{P}_r)\boldsymbol{X}_{-i}\boldsymbol{L}_{i-1}^{-1}\boldsymbol{X}_{-i}'\boldsymbol{X}_r\right). \tag{32}$$

But $\boldsymbol{X}_{-i}\boldsymbol{L}_{i-1}^{-1}\boldsymbol{U}_{i-1} = \boldsymbol{X}_{-i}\boldsymbol{B}_{i-1} = \boldsymbol{0}$ by the induction hypothesis. Moreover, since by the induction hypothesis $\boldsymbol{L}_{i-1}^{-1}$ is a generalized inverse of $\boldsymbol{Q}_{i-1}$, it is well known that $\boldsymbol{X}_{-i}\boldsymbol{L}_{i-1}^{-1}\boldsymbol{X}_{-i}' = \boldsymbol{P}_{1\cdots(i-1)}$ (e.g. Christensen 1996, Theorem B.44). Thus, using assumption (20) and Lemma 1, the second component of the block matrix on the right of (32) is zero, and the proof of the induction step is complete.

**Proof of Theorem 5.** Using the lower/upper triangular decomposition $\boldsymbol{Q}_{\tau_0} = \boldsymbol{L}_{\tau_0} - \boldsymbol{U}_{\tau_0}$ again, $\boldsymbol{B}_{\tau_0} = \boldsymbol{L}_{\tau_0}^{-1}\boldsymbol{U}_{\tau_0}$. But $\boldsymbol{Q}_{zz}$ is block diagonal by the assumptions on $\boldsymbol{Z}$, $\boldsymbol{M}(\boldsymbol{\tau})$ and $\boldsymbol{M}_{r+1}$, so

$$\boldsymbol{L}_{\tau_0} = \begin{pmatrix} \boldsymbol{L}_{xx,\tau_0} & \boldsymbol{0} \\ \boldsymbol{Q}_{zx} & \boldsymbol{Q}_{zz} \end{pmatrix} \text{ and } \boldsymbol{U}_{\tau_0} = \begin{pmatrix} \boldsymbol{U}_{xx,\tau_0} & -\boldsymbol{Q}_{xz} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Equation (24) is immediate using (31).

In part (b), $\lim_{\tau_0\to\infty} \boldsymbol{L}_{xx,\tau_0}^{-1} = \boldsymbol{0}$ but $\lim_{\tau_0\to\infty} \boldsymbol{B}_{xx,\tau_0} = \boldsymbol{B}_{xx,\infty}$ exists. Equation (25) follows immediately from (24).

Finally, for part (c), $\lim_{\tau_0\to 0} \boldsymbol{L}_{xx,\tau_0} = \boldsymbol{M}(\boldsymbol{\tau})$ and $\lim_{\tau_0\to 0} \boldsymbol{U}_{xx,\tau_0} = \boldsymbol{0}$, so $\lim_{\tau_0\to 0} \boldsymbol{B}_{xx,\tau_0} = \boldsymbol{0}$. Substituting these values in (24) gives the limit (26).

Figure 1: BIB Design Example

Figure 2: Logistic Example